CLAIMS

We claim:

1. A computerized method of producing a mechanism model based on features and responses of a set of data objects, said computerized method comprising:

establishing a description of each data object, based on a comparison between a set of features of the data objects and a set of descriptors;

selecting a group of data objects that have similar descriptions and that cooperatively exhibit a particular response characteristic;

establishing said mechanism model based on commonality of features among the data objects in the selected group; and

outputting data indicative of said mechanism model.

2. A computerized method of producing a mechanism model based on features and response characteristics of a set of data objects, said computerized method comprising:

establishing a description of each data object, based on a comparison between a set of features of the data objects and a set of descriptors;

selecting at least one group of data objects that have similar descriptions and that cooperatively exhibit a particular response characteristic, said group of data objects having a set of discriminating features defining similarity of the data objects in said group;

identifying at least one common subset of features of the data objects in said group based at least in part on a measure of how much said at least one common subset of features participated in defining the discriminating features of said group; and

outputting said at least one common subset of features as a mechanism model.

3. A computerized method as claimed in claim 2, wherein selecting at least one group of data objects that have similar descriptions and that cooperatively exhibit a particular response characteristic comprises:

grouping the data objects based on similarity of their respective descriptions, so as to produce groups of data objects; and

selecting at least one of said groups of data objects that contains data objects having said particular response characteristic.

4. A computerized method as claimed in claim 3, wherein the selected group contains at least a predetermined number of data objects having said particular response characteristic.

5. A computerized method as claimed in claim 3, wherein the selected group contains at least a predetermined percent of data objects with said particular response characteristic.

6. A computerized method as claimed in claim 3, wherein grouping the data objects based on similarity of their respective descriptions comprises clustering said data objects.

7. A computerized method as claimed in claim 3, wherein clustering said data objects comprises applying a self-organizing-map keyed to said descriptors.

8. A computerized method as claimed in claim 2, wherein selecting at least one group of data objects comprises grouping the data objects based on both their respective descriptions and their respective response characteristics.

9. A computerized method as claimed in claim 8, wherein grouping the data objects based on both their respective descriptions and their respective response characteristics comprises grouping the data objects based on features that the data objects have in common and along a dimension related to the response characteristics of the data objects.

10. A computerized method of producing a mechanism model based on features and response characteristics of a set of data objects, said computerized method comprising:

(a) iteratively performing at least the following steps:

(i) establishing a description of each data object, based on a comparison between a set of features of the data objects and a set of descriptors,

(ii) selecting a group of data objects that have similar descriptions and that cooperatively exhibit a particular response characteristic,

(iii) establishing a new descriptor based on commonality of features among the data objects in the selected group, and

(iv) adding said new descriptor to said set of descriptors; and

(b) outputting data indicative of at least one new descriptor learned in step (iv), whereby said new descriptor defines a mechanism model.

11. A computerized method of producing a mechanism model based on features and response characteristics of a set of data objects, said computerized method comprising:

(a) iteratively performing at least the following steps:

(i) establishing a description of each data object, based on a comparison between a set of features of the data objects and a set of descriptors,

(ii) selecting at least one group of data objects that have similar descriptions and that cooperatively exhibit a particular response characteristic, said group of data objects having a set of discriminating features defining similarity of the data objects in said group,

(iii) identifying at least one common subset of features of the data objects in said group based at least in part on a measure of how much said at least one common subset of features participated in defining the discriminating features of said group;

(iv) adding said at least one common subset of features to said set of descriptors as a new descriptor; and

(b) outputting data indicative of at least one common subset of features identified in step (iii), whereby said at least one common subset of features defines a mechanism model.

12. A computerized method as claimed in claim 11, wherein selecting at least one group of data objects that have similar descriptions and that cooperatively exhibit a particular response characteristic comprises:

grouping the data objects based on similarity of their respective descriptions, so as to produce groups of data objects; and

selecting at least one of said groups of data objects that contains data objects having said particular response characteristic.

13. A computerized method as claimed in claim 12, wherein the selected group contains at least a predetermined number of data objects having said particular response characteristic.

14. A computerized method as claimed in claim 12, wherein the selected group contains at least a predetermined percent of data objects having said particular response characteristic.

15. A computerized method as claimed in claim 12, wherein grouping the data objects based on similarity of their respective descriptions comprises clustering said data objects.

16. A computerized method as claimed in claim 12, wherein clustering said data objects comprises applying a self-organizing-map keyed to said descriptors.

17. A computerized method as claimed in claim 11, wherein selecting at least one group of data objects comprises grouping the data objects based on both their respective descriptions and their respective response characteristics.

18. A computerized method for generating a mechanism model defining a feature set likely to give rise to a specified response, said computerized method comprising, in combination:

(a) receiving a data set representing a plurality of objects, each object defining a set of features and a response characteristic;

(b) characterizing each object by a feature vector based on a comparison of a set of reference descriptors with the set of features defined by the object;

(c) clustering the objects based on their feature vectors and thereby producing a set of clusters, each cluster containing one or more objects, and each cluster defining a feature template associated with the features of the one or more objects in the cluster;

(d) selecting a hot spot of clustered objects based at least in part on a concentration of a specified response characteristic among the clustered objects, the hot spot having a set of discriminating features defining similarity among the objects in the hot spot;

(e) mapping the discriminating features of the hot spot to the objects in the hot spot so as to discover a feature set in the hot spot that is common among objects in the hot spot, said feature set defining said mechanism model; and

(f) outputting a data set indicating said mechanism model.

19. A method as claimed in claim 18, further comprising outputting a data set indicating the feature templates of said clusters.

20. A computerized method for producing a mechanism model representing a feature set likely to correspond to a specified response characteristic, said computerized method comprising, in combination:

(a) receiving a data set representing a plurality of data objects, each of said data objects having a set of features and a response characteristic;

(b) assembling a set of descriptors each of which defines a feature set comprising one or more features;

(c) performing the following routine at least twice, with respect to at least a plurality of said data objects:

(i) establishing a vector for each of said data objects, wherein each element of the vector for a data object indicates the presence or absence in said data object of a respective one of the descriptors of said set of descriptors,

(ii) clustering said data objects according to their vectors, so as to establish a set of clusters each containing one or more of said data objects, each of said clusters having a cluster template defining features associated with the one or more data objects in the cluster,

(iii) identifying a group of clustered data objects having at least a threshold concentration of a specified response characteristic, said group of clustered data objects having a set of discriminating features defining similarity among the clustered data objects in said group,

(iv) identifying a subset of said discriminating features that is common to all of the clustered data objects in said group, and

(v) adding said subset of discriminating features as a new descriptor to said set of descriptors; and

(d) outputting a data set indicative of at least one new descriptor, wherein said at least one new descriptor represents said mechanism model.

21. A computerized method of determining a response characteristic of a test data object based on an analysis of a set of training data objects, said test data object defining features, said computerized method comprising in order:

(a) iteratively performing at least the following steps with respect to said set of training data objects, each of said training data objects defining features and a response characteristic:

(i) establishing a descriptor vector for each of said training data objects based on a comparison of features of the training data objects with a set of descriptors,

(ii) clustering the training data objects according to their descriptor vectors, so as to establish a set of clusters, each of the clusters containing one or more training data objects, and each of the clusters having a cluster template defining a weighted set of descriptors, wherein the descriptor vectors of the training data objects in a given cluster most closely match the template of the given cluster compared to the templates of other clusters,

(iii) selecting a group of the clustered training data objects based on a concentration of a particular response characteristic among the clustered training data objects,

(iv) learning a new descriptor based on commonality of features among the clustered training data objects of the selected group, and

(v) adding said new descriptor to said set of descriptors;

(b) establishing a test descriptor vector for said test data object based on a comparison of features of the test data object with the set of descriptors;

(c) selecting as a test cluster the cluster having a template that most closely matches said test descriptor vector;

(d) determining a representative response characteristic of the training data objects in said test cluster; and

(e) concluding that the response characteristic of said test data object is likely to be said representative response characteristic.

22. A computerized method of producing a pharmacophore based on structural features and activity characteristics of a set of molecules, said computerized method comprising:

establishing a description of each molecule, based on a comparison between a set of structural features of the molecules and a set of descriptors;

selecting a group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic;

establishing said pharmacophore based on commonality of structural features among the molecules in the selected group; and

outputting data indicative of said pharmacophore.

23. A computerized method of producing a pharmacophore based on structural features and activity characteristics of a set of molecules, said computerized method comprising:

establishing a description of each molecule, based on a comparison between a set of structural features of the molecules and a set of descriptors;

selecting at least one group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic, said group of molecules having a set of discriminating features defining similarity of the molecules in said group;

identifying at least one common subset of features of the molecules in said group based at least in part on a measure of how much said at least one common subset of features participated in defining the discriminating features of said group; and

outputting said at least one common subset of features as a pharmacophore.

24. A computerized method as claimed in claim 23, wherein selecting at least one group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic comprises:

grouping the molecules based on similarity of their respective descriptions, so as to produce groups of molecules; and

selecting at least one of said groups of molecules that contains molecules having said particular activity characteristic.

25. A computerized method as claimed in claim 24, wherein the selected group contains at least a predetermined number of molecules having said particular activity characteristic.

26. A computerized method as claimed in claim 24, wherein the selected group contains at least a predetermined percent of molecules with said particular activity characteristic.

27. A computerized method as claimed in claim 24, wherein grouping the molecules based on similarity of their respective descriptions comprises clustering data representing said molecules.

28. A computerized method as claimed in claim 24, wherein clustering said data comprises applying a self-organizing-map keyed to said descriptors.

29. A computerized method as claimed in claim 23, wherein selecting at least one group of molecules comprises grouping the molecules based on both their respective descriptions and their respective activity characteristics.

30. A computerized method as claimed in claim 29, wherein grouping the molecules based on both their respective descriptions and their respective activity characteristics comprises grouping the molecules based on structural features that the molecules have in common and along a dimension related to the activity characteristics of the molecules.

31. A computerized method of producing a pharmacophore based on features and activity characteristics of a set of molecules, said computerized method comprising:

(a) iteratively performing at least the following steps:

(i) establishing a description of each molecule based on a comparison between a set of structural features of the molecules and a set of descriptors,

(ii) selecting a group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic,

(iii) establishing a new descriptor based on commonality of structural features among the molecules in the selected group, and

(iv) adding said new descriptor to said set of descriptors; and

(b) outputting data indicative of at least one new descriptor learned in step (iv), whereby said new descriptor defines a pharmacophore.

32. A computerized method of producing a pharmacophore based on features and activity characteristics of a set of molecules, said computerized method comprising:

(a) iteratively performing at least the following steps:

(i) establishing a description of each molecules, based on a comparison between a set of structural features of the molecules and a set of descriptors,

(ii) selecting at least one group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic, said group of molecules having a set of discriminating features defining similarity of the molecules in said group,

(iii) identifying at least one common subset of features of the molecules in said group based at least in part on a measure of how much said at least one common subset of features participated in defining the discriminating features of said group;

(iv) adding said at least one common subset of features to said set of descriptors as a new descriptor; and

(b) outputting data indicative of at least one common subset of features identified in step (iii), whereby said at least one common subset of features defines a pharmacophore.

33. A computerized method as claimed in claim 32, wherein selecting at least one group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic comprises:

grouping the molecules based on similarity of their respective descriptions, so as to produce groups of molecules; and

selecting at least one of said groups of molecules that contains molecules having said particular activity characteristic.

34. A computerized method as claimed in claim 33, wherein the selected group contains at least a predetermined number of molecules having said particular activity characteristic.

35. A computerized method as claimed in claim 33, wherein the selected group contains at least a predetermined percent of molecules having said particular activity characteristic.

36. A computerized method as claimed in claim 33, wherein grouping the molecules based on similarity of their respective descriptions comprises clustering said molecules.

37. A computerized method as claimed in claim 33, wherein clustering said molecules comprises applying a self-organizing-map keyed to said descriptors.

38. A computerized method as claimed in claim 32, wherein selecting at least one group of molecules comprises grouping the molecules based on both their respective descriptions and their respective activity characteristics.

39. A computerized method for producing a pharmacophore representing a chemical structure likely to have a specified activity, said computerized method comprising, in combination:

(a) assembling a set of molecule data strings each representing a molecule, and assembling activity data indicative of an activity characteristic for each of said molecules;

(b) assembling a set of descriptor data strings each representing a descriptor of a chemical structure that may be present or absent in one of said molecules;

(c) performing the following routine at least twice, with respect to at least a plurality of said molecules:

(i) for each of said molecules, establishing a vector indicating for each of said chemical descriptors whether the chemical structure represented by the descriptor data string is present in the molecule represented by the molecule data string,

(ii) clustering said molecules according to their vectors, so as to establish a set of clusters based on structural similarity of the molecules, each of the clusters thus representing one or more of the molecules, and each of the clusters having a cluster template defining a weighted set of descriptors, wherein the vectors of the molecules represented by a given cluster most closely match the template of the given cluster compared to the templates of other clusters,

(iii) identifying a group of clustered molecules having at least a threshold concentration of a specified activity characteristic, said group of clustered molecules having a set of discriminating structural features defining similarity among the clustered molecules in said group,

(iv) identifying a subset of said discriminating structural features that is common to all of the clustered molecules in said group, said subset of discriminating structural features defining a new descriptor, and

(v) adding to said set of descriptor data strings a new data string representing said new descriptor; and

(d) outputting a data set indicative of at least one new descriptor, wherein said at least one new descriptor represents said pharmacophore.

40. A computerized method as claimed in claim 39, wherein assembling a set of molecule data strings each representing a molecule and assembling activity data indicative of an activity characteristic for each of said molecules comprises receiving a single set of data including both said molecule data strings and said activity data per molecule.

41. A computerized method as claimed in claim 39, wherein assembling a set of molecule data strings each representing a molecule and assembling activity data indicative of an activity characteristic for each of said molecules comprises receiving a first data set comprising said molecule data strings and a second data set comprising said activity data.

42. A computerized method as claimed in claim 39, wherein establishing a vector for a molecule comprises querying each of said descriptor data strings against the molecule data string and responsively recording in said vector whether each respective descriptor data string is present in said molecule data string.

43. A computerized method as claimed in claim 39, wherein clustering said molecules according to their vectors comprises clustering their vectors.

44. A computerized method as claimed in claim 39, wherein clustering said molecules according to their vectors comprises applying a self-organizing-map to organize data representing said molecules.

45. A computerized method as claimed in claim 44, wherein said data representing said molecules comprises said vectors.

46. A computerized method as claimed in claim 39,

wherein the activity characteristic of each molecule is active or inactive, and

wherein identifying a group of clustered molecules having at least a threshold concentration of a specified activity characteristic comprises identifying a group of clustered molecules in which no molecule is inactive.

47. A computerized method as claimed in claim 39,

wherein the activity characteristic of each molecule is active or inactive, and

wherein identifying a group of clustered molecules having at least a threshold concentration of a specified activity characteristic comprises identifying a cluster of molecules having more than a predetermined number of active molecules.

48. A computerized method as claimed in claim 39,

wherein clustering said molecules according to their vectors comprises applying a self-organizing-map to organize data representing said molecules, and

wherein identifying a group of clustered molecules having at least a threshold concentration of a specified activity characteristic further comprises evaluating the activity characteristic of neighboring clusters in said self-organizing-map.

49. A computerized method as claimed in claim 39,

wherein clustering said molecules according to their vectors comprises applying a self-organizing-map to organize data representing said molecules, and

wherein the set of discriminating structural features defining similarity among the clustered molecules in the group comprises the template of a cluster containing molecules in said group.

50. A computerized method as claimed in claim 39, wherein identifying a subset of said discriminating structural features that is common to all of the clustered molecules in said group comprises weighing structural components of the molecules in the group according to participation of such structural components in defining the group.

51. A computerized method as claimed in 50, wherein said structural components comprise atoms.

52. A computerized method as claimed in claim 51, wherein said structural components further comprise bonds.

53. A computerized method as claimed in claim 50, wherein weighing structural components of the molecules in the group according to participation of such structural components in defining the group comprises, for each molecule in the group, assigning a weight to each structural component of the molecule based on the number of times the structural component appears in the set of discriminating structural features of said group.

54. A computerized method as claimed in claim 53, wherein identifying a subset of said discriminating structural features that is common to all of the clustered molecules in said

group further comprises identifying a maximum common substructure among said clustered molecules.

55. A computerized method as claimed in claim 54, wherein identifying a maximum common substructure among said clustered molecules comprises applying a genetic algorithm.

56. A computerized method as claimed in claim 39, wherein clustering said molecules according to their vectors comprises applying a self-organizing-map to organize data representing said molecules, and

wherein identifying a subset of said discriminating structural features that is common to all of the clustered molecules in said group comprises selecting said subset of discriminating structural features and verifying whether said subset of discriminating structural features is also present in a neighboring cluster in said self-organizing-map.

57. A computerized method as claimed in claim 39, wherein each of said descriptors represents a chemical structure comprising a feature selected from the group consisting of an atom, an atom pair, a bond, a proton donor-acceptor pair, an aromatic ring, a shape, a size and an orientation.

58. A computerized method as claimed in claim 39, wherein each of said activity characteristics represents a single activity measurement from an assay.

59. A computerized method as claimed in claim 39, wherein each of said activity characteristics represents a plurality of activity measurements.

60. A computerized method as claimed in claim 39, further comprising, in a second or later iteration of said routine, when establishing said vector for a molecule in step (a), setting said vector to indicate absence of a chemical structure represented by a descriptor data string if the chemical structure is wholly subsumed by a new feature subset as defined by a new descriptor added in step (v).

61. A computerized method as claimed in claim 39, further comprising, in a second or later iteration of said routine, assessing performance of descriptor learned in a previous iteration.

62. A computerized method for producing a pharmacophore representing a chemical structure likely to have a specified activity, said computerized method comprising, in combination:

assembling a set of molecule data strings each representing a molecule, and assembling activity data indicative of an activity characteristic for each of said molecules;

assembling a set of descriptor data strings each representing a descriptor of a chemical structure that may be present or absent in one of said molecules;

for each of said molecules, establishing a vector indicating for each of said chemical descriptors whether the chemical structure represented by the descriptor data string is present in the molecule represented by the molecule data string;

clustering said molecules according to their vectors, so as to establish a set of clusters based on structural similarity of the molecules, each of the clusters thus representing one or more of the molecules, and each of the clusters having a cluster template defining a weighted set of descriptors, wherein the vectors of the molecules represented by a given cluster most closely match the template of the given cluster compared to the templates of other clusters;

identifying a group of clustered molecules having at least a threshold concentration of a specified activity characteristic, said group of clustered molecules having a set of discriminating structural features defining similarity among the clustered molecules in said group;

identifying a subset of said discriminating structural features that is common to all of the clustered molecules in said group, said subset of discriminating structural features defining a pharmacophore; and

outputting data indicative of said pharmacophore.

63. A computer-readable medium embodying a set of machine language instructions executable by a computer for analyzing an input set of data representing a set of molecules and thereby establishing a pharmacophore, each of said molecules having structural features and an activity characteristic, said machine language instructions comprising instructions for performing the following functions:

establishing a description of each molecule, based on a comparison between a set of structural features of the molecules and a set of descriptors;

selecting at least one group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic, said group of molecules having a set of discriminating features defining similarity of the molecules in said group;

identifying at least one common subset of features of the molecules in said group based at least in part on a measure of how much said at least one common subset of features participated in defining the discriminating features of said group; and

outputting said at least one common subset of features as a pharmacophore.

64. A processing system for modeling chemical structure-activity relationships through artificial intelligence analysis of a data set representing molecules, each of the molecules having a set of features and an activity characteristic, said processing system comprising, in combination:

means for establishing a description of each molecule, based on a comparison between a set of structural features of the molecules and a set of descriptors;

means for selecting at least one group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic, said group of molecules having a set of discriminating features defining similarity of the molecules in said group;

means for identifying at least one common subset of features of the molecules in said group based at least in part on a measure of how much said at least one common subset of features participated in defining the discriminating features of said group; and

means for outputting said at least one common subset of features as a mechanism model representing a chemical structure likely to give rise to said particular activity characteristic.

65. A computer-readable medium embodying a set of machine language instructions executable by a computer for analyzing an input set of data representing a set of molecules and thereby establishing a pharmacophore, each of said molecules having structural features and an activity characteristic, said machine language instructions comprising instructions for performing the following functions:

(a) iteratively performing at least the following steps:

(i) establishing a description of each molecules, based on a comparison between a set of structural features of the molecules and a set of descriptors,

(ii) selecting at least one group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic, said group of molecules having a set of discriminating features defining similarity of the molecules in said group,

(iii) identifying at least one common subset of features of the molecules in said group based at least in part on a measure of how much said at least one common subset of features participated in defining the discriminating features of said group;

(iv) adding said at least one common subset of features to said set of descriptors as a new descriptor; and

(b) outputting data indicative of at least one common subset of features identified in step (iii), whereby said at least one common subset of features defines a pharmacophore.

66. A processing system for modeling chemical structure-activity relationships through artificial intelligence analysis of a data set representing molecules, each of the molecules having a set of features and an activity characteristic, said processing system comprising, in combination:

(a) means for iteratively performing at least the following steps:

(i) establishing a description of each molecules, based on a comparison between a set of structural features of the molecules and a set of descriptors,

(ii) selecting at least one group of molecules that have similar descriptions and that cooperatively represent a particular activity characteristic, said group of molecules having a set of discriminating features defining similarity of the molecules in said group,

(iii) identifying at least one common subset of features of the molecules in said group based at least in part on a measure of how much said at least one common subset of features participated in defining the discriminating features of said group;

(iv) adding said at least one common subset of features to said set of descriptors as a new descriptor; and

(b) means for outputting data indicative of at least one common subset of features identified in step (iii), whereby said at least one common subset of features defines a mechanism

model representing a chemical structure likely to give rise to said particular activity characteristic.

67. A computer-readable medium embodying a set of machine language instructions executable by a computer for analyzing an input set of data representing a set of molecules and thereby establishing a pharmacophore representing a chemical structure likely to have a specified activity, each of said molecules having structural features and an activity characteristic, said machine language instructions comprising instructions for performing the following functions:

(a) assembling a set of molecule data strings each representing a molecule, and assembling activity data indicative of an activity characteristic for each of said molecules;

(b) assembling a set of descriptor data strings each representing a descriptor of a chemical structure that may be present or absent in one of said molecules;

(c) performing the following routine at least twice, with respect to at least a plurality of said molecules:

(i) for each of said molecules, establishing a vector indicating for each of said chemical descriptors whether the chemical structure represented by the descriptor data string is present in the molecule represented by the molecule data string,

(ii) clustering said molecules according to their vectors, so as to establish a set of clusters based on structural similarity of the molecules, each of the clusters thus representing one or more of the molecules, and each of the clusters having a cluster template defining a weighted set of descriptors, wherein the vectors of the molecules represented by a given cluster most closely match the template of the given cluster compared to the templates of other clusters,

(iii) identifying a group of clustered molecules having at least a threshold concentration of a specified activity characteristic, said group of clustered molecules having a set of discriminating structural features defining similarity among the clustered molecules in said group,

(iv) identifying a subset of said discriminating structural features that is common to all of the clustered molecules in said group, said subset of discriminating structural features defining a new descriptor, and

(v) adding to said set of descriptor data strings a new data string representing said new descriptor; and

(d) outputting a data set indicative of at least one new descriptor, wherein said at least one new descriptor represents said pharmacophore.

68. A processing system for modeling chemical structure-activity relationships through artificial intelligence analysis of a data set representing molecules, each of the molecules having a set of features and an activity characteristic, said processing system comprising, in combination:

(a) means for assembling a set of molecule data strings each representing a molecule, and assembling activity data indicative of an activity characteristic for each of said molecules;

(b) means for assembling a set of descriptor data strings each representing a descriptor of a chemical structure that may be present or absent in one of said molecules;

(c) a set of machine language instructions executable by a processor for performing the following functions at least twice, with respect to at least a plurality of said molecules:

(i) for each of said molecules, establishing a vector indicating for each of said chemical descriptors whether the chemical structure represented by the descriptor data string is present in the molecule represented by the molecule data string,

(ii) clustering said molecules according to their vectors, so as to establish a set of clusters based on structural similarity of the molecules, each of the clusters thus representing one or more of the molecules, and each of the clusters having a cluster template defining a weighted set of descriptors, wherein the vectors of the molecules represented by a given cluster most closely match the template of the given cluster compared to the templates of other clusters,

(iii) identifying a group of clustered molecules having at least a threshold concentration of a specified activity characteristic, said group of clustered molecules having a set of discriminating structural features defining similarity among the clustered molecules in said group,

(iv) identifying a subset of said discriminating structural features that is common to all of the clustered molecules in said group, said subset of discriminating structural features defining a new descriptor, and

(v) adding to said set of descriptor data strings a new data string representing said new descriptor; and

(d) means for outputting a data set indicative of at least one new descriptor, wherein said at least one new descriptor defines a mechanism model representing a chemical structure likely to give rise to said specified activity characteristic.